***I.J. Wireless and Microwave Technologies*, 2011, 4, 27-34**
Published Online August 2011 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijwmt.2011.04.04

Available online at http://www.mecs-press.net/ijwmt

Privacy Preserving Similarity Measurement

ZHANG Guo-rong[a,*]

[a] *Computer Teaching and Research Section, Guangzhou Academy of Fine Arts, Guangzhou, China*

Abstract

Data similarity measurement is an important direction for data mining research. This paper is concentrated on the issue of protecting the underlying attribute values when sharing data for the similarity of objects measurement and proposes a simple data transformation method: Isometric-Based Transformation (IBT). IBT selects the attribute pairs and then distorts them with Isometric Transformation. In the process of transformation, the goal is to find the proper angle ranges to satisfy the least privacy preserving requirement and then randomly choose one angle in this interval. The experiment demonstrates that the method can distort attribute values, preserve privacy information and guarantee valid similarity measurement.

Index Terms: Privacy Preserving; Similarity; Isometric Transformation

1. Introduction

Along with the development of data mining technology, a large number of private information such as shopping habits, criminal record, medical history, credit records have been widely collected and analyzed. On the one hand, these data are important for the government and business organizations in the decision-making and providing social welfare such as medical research, reduce crime, national security. On the other hand, because data mining reveals underlying pattern or all kinds of knowledge, if they are not used correctly, it may be threat to privacy and information security. Analysis of personal data may be an invasion to the personal privacy.

Data similarity measurement is an important direction for data mining research. Cluster analysis often need calculate the similarity of objects. The similarity of objects often use distance to measure, and the most commonly distance measure is Euler distance. In order to protect the initial data privacy information (such as salary, age, etc.), this paper main discusses how to achieve privacy preserving by transforming confidential digital information in the calculation of the distance between the objects.

This paper is organized as follows. Related work is reviewed in Section 2. The Isometric-Based Transformation method (IBT) is discussed in Section 3. In Section 4, we present the experimental results. Finally, Section 5 presents our conclusions.

* Corresponding author:
E-mail address: chzhzgr@163.com

2. Related Work

Privacy preserving has become an important direction for data mining research, and more and more scholars such as cryptography, statistics, or related subjects take an active part in it. At present, in order to protect the privacy of data mining, prevent the data misused, various solutions have been proposed. According to different targets, these methods are divided into two kinds: one kind is the initial data privacy protection, namely distortion, confusion, random and anonymous change initial data, because the initial data may contain confidential personal privacy information such as name, ID card number, age, wages, etc. Another kind is distributed data privacy protection, namely in two or more parties cooperate data mining, for some reason, the participants often unwilling to share data with others and only want to share data mining results. This kind of situation in scientific research, medical research and market dynamics research aspects is nothing new now. This requires people put forward privacy preserving data mining algorithm for participants, only get the final mining results, and besides, cannot obtain any other information.

Data perturbation represents one common approach in privacy-preserving data mining. The basic idea of data perturbation is to alter the data so that real individual data values cannot be recovered, while preserving the utility of the data for data mining and statistical summaries. Since the data doesn't reflect the real values of private data, even if a data item is linked to an individual that individual's privacy is not violated. This approach has been brought to a high art by the U.S. Census Bureau with the Public Use Micro data sets [1].

Some effort has been made to address the problem of privacy preservation in data clustering, especially in similarity measurement. For centralized data, one of the common methods is data transformation perturbation method. It makes sure the initial data cannot be calculated from the transformed data, and achieves the purpose of hiding private information. At the same time, it guarantees valid similarity measurement and clustering results. The feasibility of achieving PPC through geometric data transformation was studied in [2]. Different from some aspects from the work in [2], instead of distorting data for clustering using translations, scaling, rotations or even some combinations of these transformations, the work in [3] distort attribute pairs using rotations only to avoid misclassification of data points. It shows that successive rotations on normalized data will protect the underlying attribute values and get accurate clustering results. However, this method may not obtain appropriate rotation angle when users put forward higher privacy requirement. Furthermore, this method doesn’t discuss all isometric transformation.

This paper is concentrated on the issue of protecting the underlying attribute values when sharing data for the similarity of objects measurement and proposes a simple and effective data transformation: Isometric-Based Transformation (IBT). IBT selects the attribute pairs and then distorts them with Isometric Transformation. In the process of transformation, the goal is to find the proper angle ranges to satisfy the least privacy preserving requirement and then randomly choose one angle in this interval. Different from RBT in [3], IBT discuss all isometric transformation, use relative privacy protection degree to determine transform angle ranges, analyze some special cases in the isometric transformation, and make sure good privacy protection effect in most cases.

3. Isometric-Based Transformation Method

3.1. Isometric Transformation

Definition (Isometric Transformation). Let $T : E^n \to E^n$ be a transformation in the n-dimensional space. T is said to be an isometric transformation if it preserves distances satisfying the constraint: $|T(A) - T(B)| = |A - B|$ for all $A, B \in E^n$.

In Cartesian coordinates, translation, rotation and reflection are isometric transformation, and isometric transformation can always write:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & \mp\sin\theta \\ \sin\theta & \pm\cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{1}$$

Where (a_1, a_2) represent translation vector and $\theta \in [0,2\pi]$.

3.2. Privacy preserving measurement

To compare privacy protection degree by different transformation, this paper measures the level of security by the database security evaluation method in [4]. This measure is given by $S = Var(X - X')/Var(X)$ where X represents a single original attribute and X' the distorted attribute, and $Var(X) = Var(x_1, x_2, \cdots, x_N) = \frac{1}{N} \times \sum_{i=1}^{N} (x_i - \overline{x})^2$

Where $\overline{x}$ is the arithmetic mean of the values $x_1, x_2, \cdots x_N$.

Clearly, the above measure to quantify privacy is based on how closely the original values of a modified attribute can be estimated. The attributes in a data matrix have been z-score normalization before the implementation of IBT. Therefore, for all attributes, the normalized attribute value X satisfies

$$Var(X) = \frac{1}{N} \times \sum_{i=1}^{N} (x_i - \overline{x})^2 = \frac{N-1}{N} \tag{2}$$

Therefore, the measurement of privacy protection degree is given as follows:

$$P = Var(X - X') \tag{3}$$

In the implementation of IBT process, we still must use the following concepts:

The optimal privacy protection degree is $P_0 = f_{\max}(\theta)$ when $f(\theta) = \min(Var(X - X'), Var(Y - Y'))$

The relative privacy protection degree S is percentage of desire privacy protection degree and optimal privacy protection degree. It is provided by the operator before the program is run.

The minimum privacy protection degree ρ is product optimal privacy protection degree and relative privacy protection degree.

3.3. Some Theorem

Theorem 1: Translation transformation failed to protect privacy.

Proof: Using (3), privacy protection degree is given as follows:

$$Var(A_i - A_i') = 0 \qquad Var(A_j - A_j') = 0$$

This suggests that translation transformation failed to protect privacy. Therefore, (1) can be written as:

T_1:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = T_1 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{4}$$

$T_{2:}$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = T_2 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{5}$$

Where $\theta \in [0,2\pi]$

Theorem 2: Let attribute vector pairs V(A$_i$, A$_j$), (i≠j) in data matrix $D_{m\times n}$ implement the transformation T$_1$, Attribute pairs achieve optimal privacy protection degree when $\theta = \pi$, and optimal privacy protection degree is given by $P_0 = 4\dfrac{m-1}{m}$.

Proof: Let $A_i = (a_{1i}, a_{2i}, \cdots a_{mi})$, $A_j = (a_{1j}, a_{2j}, \cdots a_{mj})$, privacy protection degree is given as follows:

$$g(\theta) = Var(X - X') + Var(Y - Y') = 4(1 - \cos\theta)\frac{m-1}{m}$$

When $\theta = \pi$, $g(\theta)$ reaches its maximum, and $Var(X - X') = Var(Y - Y')$.

Therefore, $P_0 = f_{\max}(\theta) = {g(\pi)}/{2} = 4\dfrac{m-1}{m}$.

Theorem 3: Let attribute vector pairs V(A$_i$, A$_j$), (i≠j), in data matrix $D_{m\times n}$ implements the transformation T$_2$, Attribute pairs achieve optimal privacy protection degree when $\theta = \pi/2$ or $\theta = 3\pi/2$, and optimal privacy protection degree is given by $P_0 = 2(\dfrac{m-1}{m}) + \dfrac{2}{m}\left|\sum_{k=1}^{m} a_{ki}a_{kj}\right|$.

Proof: Similarly, following Theorem 2 proof method, privacy protection degree is given as follows:

$$g(\theta) = 4\frac{m-1}{m} - 4\frac{\sin\theta}{n}\sum_{k=1}^{m} a_{ki}a_{kj}$$

When $\theta = \pi/2$ or $\theta = 3\pi/2$, $g(\theta)$ reaches its maximum, and $Var(X - X') = Var(Y - Y')$.

Therefore, $P_0 = f_{\max}(\theta) = {g_{\max}(\theta)}/{2} = 2\dfrac{m-1}{m} + \dfrac{2}{m}\left|\sum_{k=1}^{m} a_{ki}a_{kj}\right|$

Theorem 4: In the transformation process, the distorted attribute pairs maybe satisfy the least privacy preserving requirement when $\theta = 0, \pi/2, \pi, 3\pi/2$. But intuitively, the privacy is easy to be violated.

Proof: Isometric transformationT$_1$, T$_2$ take the following values respectively when $\theta = 0, \pi/2, \pi, 3\pi/2$:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$$

These transformations or no effect, or simply exchanged attributes on position, or only added minus. Although they satisfy the least privacy preserving requirement, intuitively, the privacy still is easy to be violated. Therefore, in the process of transformation, it should avoid select these values angle.

3.4. Isometric-Based Transformation Method

IBT selects the attribute pairs and then distorts them with Isometric Transformation. The goal is to find the proper angle ranges to satisfy the least privacy preserving requirement and then randomly choose one angle in this interval. The procedure to distort the attributes of a data matrix has essentially 2 major steps, as follows:

Step 1. Selecting the attribute pairs: We select k pairs of attributes A$_i$ and A$_j$ in D, where i≠j. If n is odd, the last attribute selected is distorted along with any other attribute distorted.

Step 2. Distorting the attribute pairs: The pairs of attributes selected previously are distorted as follows:

- Randomly selecting a kind of isometric transformation, use the (4) or (5) transform attribute pairs, get the new attribute pairs that contains transform angle variables.
- Computing optimal privacy protection degree P$_0$, By Theorem 2 and Theorem 3 can quickly calculate $P_0 = f_{\max}(\theta)$.
- Computing the minimum privacy protection degree ρ, ρ=P$_0$ * S, where S the requirement of the relative privacy protection degree.
- Computing the proper θ ranges to satisfy the minimum requirement of privacy protection degree, θ must satisfy the following conditions:

$$Var(A_i - A_i') > \rho \text{ , } Var(A_j - A_j') > \rho$$

- Randomly selecting θ in the ranges, but $\theta \neq 0, \pi/2, \pi, 3\pi/2$
- Output the transformed attribute pairs A'$_i$, A'$_j$, (i≠j).

In the transformation process, if attribute pairs consists of an original attribute and another attribute distorted, the privacy protection degree must be calculated by their original attribute values. This will prevent attribute distorted returning to the original state in the second transformation.

The inputs for the IBT algorithm are a normalized data matrix D and a set of k the relative privacy protection degree S$_k$. We assume that there are k pairs of attributes to be distorted. The output is the transformed data matrix D which is shared for similarity measurement. The sketch of the IBT algorithm is given as follows:

```
IBT Algorithm
Input: $D_{m\times n}, S_k$
Output: $D'_{m\times n}$
```

```
1. $k \leftarrow [{}^{n}\!/_{2}]$
2. $P_k \leftarrow k$ pairs $(A_i, A_j)$ in $D$ such that $1 \le i, j \le n$ and $i \ne j$
3.For each selected pair $P_k$ do
3.1. Randomly selecting a kind of isometric transformation
3.1.1 Transform $T_1$ : $V(A_i', A_j') \leftarrow T_1 \times V(A_i, A_j)$
3.1.2 Transform $T_2$ : $V(A_i', A_j') \leftarrow T_2 \times V(A_i, A_j)$
3.2.Computing $P_0 = f_{\max}(\theta)$ , where
 $f(\theta) = \min(Var(A_i - A_i'), Var(A_j - A_j'))$
3.3.Computing $\rho_k = P_0 * S_k$
3.4.Computing $Var(A_i - A_i') > \rho_k$ , $Var(A_j - A_j') > \rho_k$
3.5. $\theta_k \leftarrow \theta$ , $\theta$ satisfy $Var(A_i - A_i') > \rho_k$ and $Var(A_j - A_j') > \rho_k$
3.6. $V(A_i', A_j') \leftarrow T_1 \times V(A_i, A_j)$ , $V(A_i', A_j') \leftarrow T_2 \times V(A_i, A_j)$
// Output the distorted attributes of $D'_{m \times n}$
End for
End Algorithm
```

4. Experiment

4.1. Experimental Results

In this section, we present a simple experiment about some records (Table 1 and Table 2). These records contain real data of the Cardiac Arrhythmia Database available at the UCI Repository of Machine Learning Databases [3,5]. We select the pairs of attributes [age; heart_rate] to distort, and assume that relative privacy protection degree is 0.5. Fig. 1 shows every point rotates 2.6 CCW by transformation T_1, and the pairs of attributes get a very good disturbed, and the distance between every two points is remained. Fig. 2 shows every point symmetry transform by transformation T_2, and the distance between every two points is also remained. The experiment demonstrates that the method efficiently distorts attribute values, preserves privacy information and guarantees valid similarity measurement.

Table 1. Some records of the cardiac arrhythmia database

ID	age	weigh	heart_rate
1	75	80	63
2	56	64	53
3	40	52	70
4	28	58	76
5	44	90	68

Table 2. The corresponding normalized database

ID	age	weigh	heart_rate
1	1.4809	0.7095	-0.3476
2	0.4151	-0.3041	-1.5061
3	-0.4824	-1.0642	0.4634
4	-1.1556	-0.6841	1.1586
5	-0.2580	1.3430	0.2317

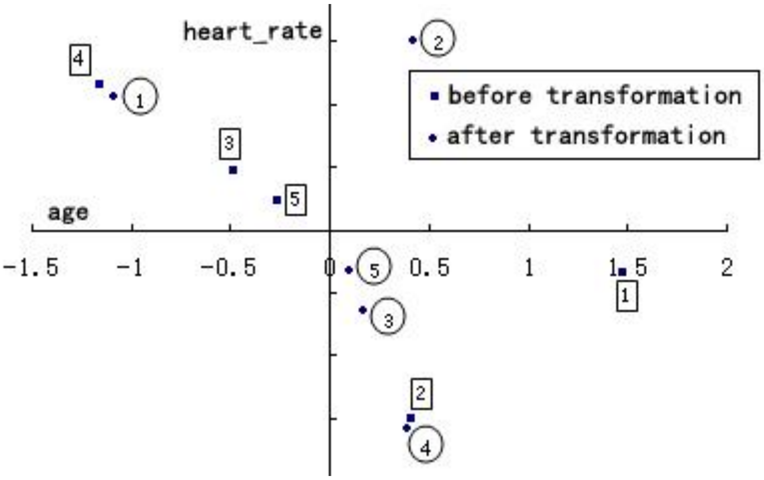

Fig. 1. Position variation by T_1

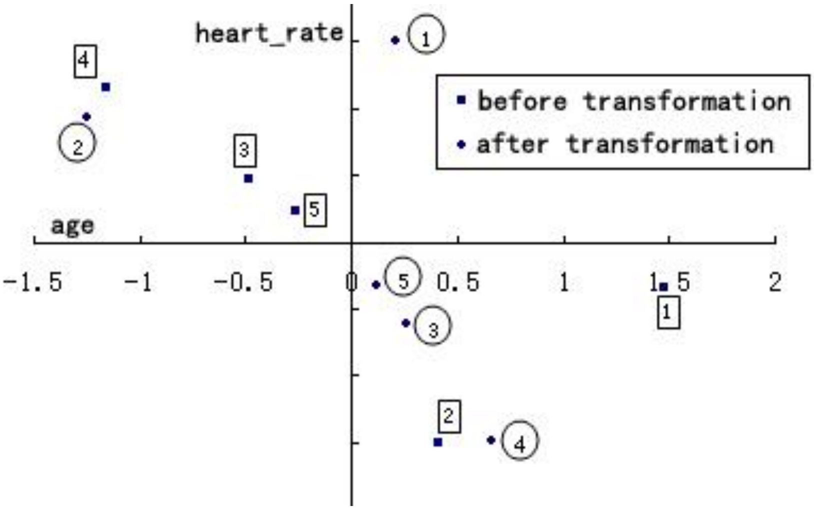

Fig. 2 Position variation by T_2

5. Conclusions

In this paper, we propose a method called Isometric-Based Transformation (IBT). IBT selects the attribute pairs and then distorts them with Isometric Transformation. In the process of transformation, the goal is to find the proper angle ranges to satisfy the least privacy preserving requirement and then randomly choose one angle in this interval. The experiment demonstrates that the method efficiently distorts attribute values, preserves privacy information and guarantees valid similarity measurement.

Our contributions in this paper can be summarized as follows: First, we discuss all isometric transformation, and can randomly select a kind of isometric transformation in the process of transformation. Second, we use relative privacy protection degree to determine transform angle ranges, and IBT can obtain appropriate rotation angle when users put forward higher privacy requirement. In addition, we analyze some special cases in the isometric transformation, and improve privacy protection degrees measurement method when number of attribute is odd, and make sure good privacy protection effect in most cases.

References

[1] Richard A. Moore, Jr. Controlled data-swapping techniques for masking public use microdata sets. Statistical Research Division Report Series RR 96-04, U.S. Bureau of the Census, Washington, DC, 1996.

[2] S.R.M. Oliveira, O.R. Zaïane. Privacy Preserving Clustering By Data Transformation. In Proceedings of the 18th Brazilian Symposium on Databases, Manaus, Amazonas, Brazil, October 2003, pp.304-318.

[3] S.R.M. Oliveira, O.R.Zaïane. Achieving Privacy Preservation When Sharing Data For Clustering. In Proceedings of the International Workshop on Secure Data Management in a Connected World (SDM'04) in conjunction with VLDB 2004, Toronto, Canada, August, 2004

[4] K.Muralidhar, R.Parsa, R.Sarathy. A General Additive Data Perturbation Method for Database Security. Management Science, 1999, October, 45(10):1399–1415.

[5] C.L. Blake and C.J. Merz. UCI Repository of Machine Learning Databases, University of California, Irvine, Dept. of Information and Computer Sciences, 1998